CGIAR Research Program on
Climate Change, Agriculture and Food Security (CCAFS)

Data Quality Checking
Video Transcript

October 2013

These materials were produced for and with funding from the Climate Change Agriculture and Food Security Research Program of the Consultative Group on International Agricultural Research (CGIAR).

Introduction

Data quality checking is the process of reviewing data to discover inconsistencies and anomalies and performing data cleaning to improve the quality.

Systematically checking data and eliminating errors increases the reliability of, and adds credibility, to your results.

This short video goes onto highlight key stages in an activity at which data checks should be carried out, and activities performed, to ensure the activity collects and analyses quality data.

Planning Stage – Designing Tools

Before even collecting data you can start thinking about data quality, for example how to phrase questions in a survey to minimise different interpretations.

At this stage you can also start thinking and producing a list of checks to run on the data once it has been collected.

Data Collection

Enumerators and field supervisors should be responsible for running basic data checks in the field. Identifying issues at this stage maximises the chances for incorrect data being corrected, as the enumerator can go back to the household, individual, or experimental plot.

Data Entry

If using a customised data entry system a number of basic consistency checks can be incorporated into data entry software to reduce mistakes occurring during data entry.

We generally recommend double data entry, so two individuals separately entering the data, the two sets of data can then be compared and any inconsistencies can be identified and updated.

Once the data has been entered you can run further checks by producing simple frequency tables, cross tabulations and summaries. For numeric variables, such as yields, examining statistics such as the minimum and maximum, and plotting the data can be a very useful method of identifying incorrect values, or outliers.

Transition from Raw to Primary Data

At this stage you should focus your checks on any derived variables (these are new variables you create based on the raw data, such as calculating the yield in tonnes per hectare from the original yield data and plot size), and any data manipulations that have been conducted.

Quality Checks between Stages

Data Collection to Data Entry

If the activity involves paper questionnaires or field reports checks need to be put in place to ensure all of the questionnaires/reports are handed into the office. A simple logging-in process can be implemented to help with this.

For electronic data checks need to be conducted to ensure all of the files have been transferred and are not corrupted, checks should also be made for duplicates. You should consider a backup system for your hand-held devices.

Irrespective of the data collections tools being used you must check:

- Everything from the field has been received in the office
- The data is uncorrupted and complete
- The expected number of questionnaires/electronic files have been received

Audit Trail, Backups and Versions

You should start keeping an audit trail as soon as you formally start checking your data. This can be in the form of a document containing details of the checks that took place in the field and in data entry, once the data has been entered it should include details of any corrections/changes to the data, and how any outliers were dealt with. This document should be regularly updated until the final primary dataset or datasets have been finalised.

Once you have entered your data you need to ensure you regularly start keeping backups of your data. You may be making several small changes to the dataset, you need to ensure that a record is kept of these changes and you know at all times which dataset contains the corrections and which contains the original values. If you keep regular backups then if an incorrect update is made then you can go always go back to the previous version.

During an activity you are going to have several versions of the data, the raw data, the primary data and the several datasets in-between these two stages. It is imperative that you keep track of these versions and record the changes made from one version to the next in the audit trail document.

More details about the Data Quality Checking concepts introduced in this video are available in the corresponding guide as part of the Data Management Support Pack.

Appendix I – CCAFS Data Management Support Pack

This document is part of the CCAFS Data Management Support Pack produced by the Statistical Services Centre, University of Reading, UK. The following materials are available in the pack:

0. Data Management Strategy
    a. CCAFS Data Management Strategy
1. Research Protocols
    a. Writing Research Protocols – a statistical perspective
    b. Preparation of Research Protocols – Good Practice Case Study
    c. What is a Research Protocol, and how to use one (Video & Transcript)
    d. Details of what a Research Protocol should contain (Video & Transcript)
2. Data Management Policies & Plans
    a. Creating a Data Management Plan
    b. Data Management Plan (Video & Transcript)
    c. Example Data Management Activity Plan
    d. Example Consent Form
3. Budgeting & Planning
    a. Budgeting & Planning for Data Management
    b. ToR Data Support Staff
    c. Budgeting & Planning (Video & Transcript)
4. Data Ownership
    a. Data Ownership and Authorship
    b. Template – Data Ownership Agreement
    c. CCAFS Data Ownership & Sharing Agreement
    d. Data Ownership & Authorship (Video & Transcript)
5. Data & Document Storage
    a. Creating and Using a DDS
    b. DDS Introduction – (Video & Transcript)
    c. DDS Organisation – (Video & Transcript)
    d. DDS Ownership – (Video & Transcript)
    e. Introduction to Dropbox – (Video & Transcript)
6. Archiving & Sharing
    a. Archiving & Sharing Data
    b. Data and Documents to Submit for Archiving – a checklist
    c. MetaData
    d. Archiving & Sharing (Video & Transcript)
    e. Metadata (Video & Transcript)
    f. CCAFS HBS Questionnaire
    g. CCAFS HHS Code Book
    h. CCAFS Training Manual for Field Supervisors

7. CCAFS Data Portals
    a. Portals for CCAFS Outputs
    b. AgTrials Summary
    c. CCAFS-Climate Summary
    d. DSpace Introduction
    e. Introduction to Dataverse (Video & Transcript)
    f. Creating a Dataverse (Video & Transcript)
    g. Dataverse Study Catalogue
    h. CCAFS Dataverse (Video & Transcript)
8. Data Quality & Organisation
    a. Data Quality Assurance
    b. Guidance for handling different types of Data
    c. Transition from Raw to Primary Data
    d. Data Quality Assurance (Video & Transcript)
    e. Guidance for handling different types of data (Video & Transcript)
    f. Transition from Raw to Primary Data (Video & Transcript)